A geometric approach to graph exploration

Fabienne Venant

A geometric approach to graph exploration

Venant F.
Lalic, Université Paris IV, France.

Graphs of lexical relations share important specific properties with graphs of social relations. We present here a method for bringing out the hierarchical organization of a graph of synonymy of French adjectives. The lexical structure can be visualized at different scales, revealing the relevant dimensions of the semantic space at each level, from the global macrostructure to the most fine-grained local organization. The method uses the 'small-world' structure of the graph. The graph is associated with a semantic space; each clique of the graph (in our case its maximal sets of adjectives that are all synonyms for one another) is a point of this space. This model accounts for theoretical studies on French adjectives. It also brings a new interesting light on the organisation of the French adjectival lexicon. The semantic space seems to play the same role for words that the geographical space does for humans. Words meet in the semantic space like people meet in the world. The tools presented here should apply to social graphs.

Keywords: lexical graph, small-world graph, exploration, clique, semantic space.

1 INTRODUCTION

The organization of lexical units in natural languages can be compared in many respects to social organizations. Words are organized in the lexicon as a complex network of evolving semantic relations. Such a system shares many important properties with complex systems of social relations. As regards hierarchical structuring, we are confronted with the same problem: words generally belong to more than one semantic class, due to the pervasive phenomenon of polysemy (words with several related meanings). Nevertheless, the existence of some form of hierarchical organization is unquestionable: some words have very wide meanings while others are more precise. It has been discovered recently that most of lexical graphs belonged to the class of "small-world graphs". This term denotes graphs where most nodes are neighbors of one another, but every node can be reached from other by a small number of hops or steps. These graphs are called "small-world" in reference to acquaintance networks: when they deal with social relation (ie when nodes represent people and edges connect people that know each other), they capture the small world phenomenon of strangers being linked by a mutual acquaintance introduced by Milgram (1967).

2. MATERIAL AND METHODS

2.1 Small-world networks

Small-world networks were defined by Watts and Strogatz (1998). They noted that graphs could be classified according to their clustering coefficient (C) and their characteristic path length (L).

The clustering coefficient C is a measure of how tightly the neighbors of a node in the graph are connected to each other. It is computed as follow:

Let p be a node, k its degree (number of its neighbors) and n the number of edges among them. The clustering coefficient at node p is $c(p) = 2n/k(k-1)$. It is easy to verify that c(p) lies between 0 and 1. It equals 0 if there is no edge linking any pair of neighbors of p, and 1 if all neighbors are connected with one another. Then the clustering coefficient C of the graph is the average of c(p) over all nodes.

In social terms, C measures how many of one's acquaintances know each other. So, we can understand why social networks have a high clustering coefficient (most of my friends are friends of each other)

The characteristic path length L is a measure of how far two nodes are situated one from the other in the graph. The distance between two vertices is the number of edges in a shortest path connecting them. The characteristic path length is the average of the distance over all pairs of nodes.

Small-world networks, as compared to other random graphs with the same number of nodes and edges, are characterized by clustering coefficients significantly higher than expected ($C >> Crandom \cong k/n$) and characteristic path length a bit lower than expected ($L > random \cong \ln(n)/\ln(k)$))

Additionally, a third property can be associated with small-world networks even though it is not required for that classification. Typically there is an over-abundance of hubs - nodes in the network with a high number of connections although most nodes are of low degree. These hubs serve as the common connections mediating the short path lengths between other edges. This property is often analyzed by

considering the degree of a randomly selected node. Networks with a greater than expected number of hubs will have a greater fraction of nodes with high degree, and consequently the degree distribution will be enriched at high degree values. Specifically, if a small-world network has a degree-distribution which can be fit with a power law distribution; it is taken as a sign that the network is small-world. The power-law distribution was first verified on the Web network, which is also a small-world graph but it also hold for social network (Newman 2001, Barabási et al. 2002). These networks are known as scale-free networks.

2.2 Semantic spaces

Ravasz and Barabási (2003) showed that high clustering coefficient with scale free topology determine an original combination of modularity and hierarchical organization. It is not a simply pyramidal organization. High clustering coefficient assures that "we have many small clusters, which are densely interconnected. These combine to form larger but less cohesive groups, which combine again to form even larger and less interconnected clusters". The scale-free structure assures that the ratio on very connected nodes to the number of nodes in the rest of the network remains constant as the network changes in size. No node can be viewed as dominating other nodes. The structure is made of groups of node, with small clusters at the bottom and very large groups at the top. Moreover, groups of nodes may overlap at any level. "This self-similar nesting of different groups or module into each other forces a strict fine structure on real network" (Ravasz and Barabási , 2003)

Such a structure can, in many cases, be related with an underlying geometrical space. For instance, the structure of a social network is clearly related with the underlying geographical space. This geographical space presents a hierarchical structure dual of this of the graph. Acquaintance relationship is highly correlated with geographical proximity. Small clusters of strongly interconnected people correspond to small areas where few people often meet (villages or district in a city). These clusters are connected one to another, and combine to form larger and larger clusters corresponding with larger areas (like cities or countries). If we consider now other type of small-world graph, we assume that there is always an underlying space which can reveal the hierarchical structure, even though most of the time the nature of this space is more abstract than a geographical map. This is the main idea of the method of exploration described in this paper: we study the structure of the French adjectival lexicon by building the semantic spaces associated with a graph of synonymy.

In this study, we work on a synonymy graph of French adjectives extracted from a general dictionary of French synonyms1. We choose to work on the French adjectival lexicon, because it is the subject of many works in linguistics. The French adjectival lexicon, even if it is almost unexplored from a computational viewpoint, is very well described from a linguistic viewpoint. Our exploration of the graph has been guided by these linguistics descriptions.

The graph studied here, called Synadj, is a graph with 3 699 vertices and 22 568 links. The average degree is 6.10 synonyms for an adjective. We can see several adjectives which are clearly hubs, ie their degrees (numbers of synonyms) are very higher than the average degree. These adjectives are for example:

beau (*lovely, beautiful, good, fine, nice*) and *bon* (*good, right, kind, …*): degree> 150

dur (*hard, tough, harsh*), *extraordinaire* (*extraordinary*), *fort (strong), grand (big, tall, long, large), mauvais (bad, wrong)* and *vif (lively, fast, sharp*): degree > 100.

We computed the clustering coefficient of Synadj C=0.28 and its characteristic path length L=4.04. For an equivalent random graph (same number of vertices n and same average degree k), we have L = log(n)/(log(k)) = 4,48 et C =k/n = 0,0017. A comparison of these values shows that Synadj has a small-world structure.

2. 3 French adjectives

Linguists usually distinguish two classes of French adjectives: qualificative adjectives and adjectives of relation:

- A qualificative adjective modifies the noun and gives information about it. It can indicate a quality like *rouge* (*red*) in *un livre rouge* (*a red book*) or *étrange* (*strange*) in *un langage étrange* (*a strange language*). This kind of adjectives gives informations about the scope of reference of a noun.
- A relationnel adjective don't determine the noun like a qualificative adjective, but it indicates the relation between this noun and an other noun, with the general meaning "of, relating to or like (the

[1] The general dictionnary of French synonyms is managed by J.L Manguin at the CRISCO research laboratory, at the University of Caen (France). It is available on th Web (http:// www.crisco.unicaen .fr)

noun)" (the precise range of meanings, and shades of meaning, varies case by case). For example *élections présidentielles* (*presidential elections*), *langage enfantin* (*baby's talk*)

In 2004, Romero proposed to distinguish a third class of French adjectives called intensive adjectives used to intensify a positive or negative property of the noun. For example *énorme* in *un succès énorme* (*a briliant success*) or *méchant* in *méchant soleil (nasty sun)*

Actually it is quite difficult to characterize a French adjective by itself. For example, the French adjective *enfantin*, usually considered as a relationnal adjective, can also be used as a qualificative adjective (*une remarque enfantine / an infantil remark*) and even as an intensive adjective (*un problème enfantin / a very easy problem*). Almost all of the French adjectives can be found in the three kinds of uses. Thus, Goes and Romero suggered that it is more relevant to talk about classes of uses than about classes of adjectives. One aim of our exploration is to characterize these classes of adjectival uses.

2.4 Method

As said above, the idea is to associate the graph with an underlying semantic space whose topological organization could reveal the structure of the graph. The small world structure of the graph incited us to use the cliques of the graph as a tool for building this semantic space[2]. A clique in a graph is a maximal set of pairwise adjacent vertices, or -in other words- an induced subgraph which is a maximal complete graph. In our case, a clique is made of adjectives which are all synonyms in a one to one relationship. By virtue of the definition, small-world networks will inevitably have high representation of cliques, and subgraphs that are a few edges shy of being cliques, i.e. small-world networks will have sub-networks that are characterized by the presence of connections between almost any two nodes within them. This follows from the requirement of a high cluster coefficient. We can consider as a first approximation that the cliques define very precise meanings that can be considered as the intersection of the meanings of all the units belonging to the clique. The main idea of the method consists in associating points of the semantic space with the cliques of the graph.

We computed all the clique of the graph Synadj[3]. Synadj has 11 900 cliques, with on average four adjectives per clique. Most of the cliques contains between 2 and 5 adjectives.

The principle of the exploration is to build semantic spaces at different scales. In order to build a local semantic space (for example associated with a given word), we select the set of relevant cliques, and compute the distances between them. We use the chi-square distance, which is well known in statistical analysis and intensely used to compute correspondences between subsets of individuals and subsets of qualitative characteristics. Then a principal component analysis is applied to reduce the dimensionality of the space.

To explore and visualize the global macrostructure of the whole French adjectival lexicon, we define a global semantic space as the Euclidian space generated by all the vertices of the graph, ie. all the adjectives of the lexicon under study. This global space will enable us to analyse the organisation of any given set of cliques among the set of all the cliques of the graph. In order to assure the correspondence between the distance in the space and the semantic proximity in the lexicon, we still use the chi-square distance

3. RESULTS

Figure 2 shows a visualization of the semantic space associated with the French adjective *sec (dry, severe, bruque…)*. It accounts for the six main meanings we can find in a dictionary.

1) Lacking water: *sable sec (dry sand).*

2) thin, bony: *un homme grand et sec (a tall and thin man).*

3) sterile, unproductive: *rester sec aux questions du professeur (cannot answer the teacher's questions).*

4) lacking sensitivity, egoist : *avoir le cœur sec (to have dry heart).*

5) brief, abrupt, lacking sweetness : *coup sec (blunt blow).*

6) alone : *atout sec (in playing cards, a singleton trumps).*

Because of the very important total number of cliques in the graph it has been impossible to visualize the global semantic space in its whole. To have an idea of its structure, we decided to explore the cloud of

[2] following the method first proposed by Ploux and Victorri (1998)

[3] The algorith used to compute the lcique can be found in Reingold et al, 1977. For a similar approach using also a graph of synonymy, see Warmesson, 1885

cliques one part at a time. We thus studied the distribution of the cliques according to their distance to the origin of the space4.

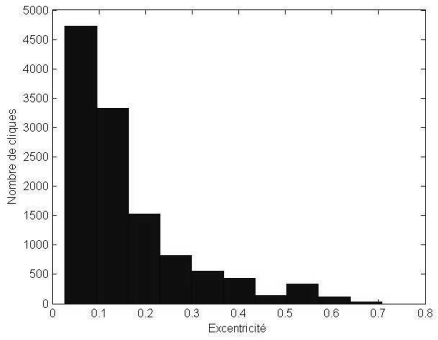

Fig 1, distribution of the cliques: number of cliques in function of the distance to the origin of the space

Most of the cliques are situated inside the central sphere (radius 0.1).This sphere still contains too much cliques for only one visualization. So we explored the sphere step by step, from the cliques near the origin to the whole sphere. Each visualisation brings new informations about the structure. We thus discovered the structure of this sphere:

The very centre of the space only contains intensive meanings like *authentique; certain; evident; incontestable* (≈ *authentic; certain; evident; incontestable*) or *agréable; charmant; enivrant; ravissant; séduisant* (≈ *agreeable; delightful; exciting; attractive*). These meanings are very general meanings and can apply to any nouns. These intensive meanings form a central core. Further from the centre the cliques become more qualificative, but still correspond to general meanings. These meanings are matter of perception, sensation or feeling. It is interesting to note that the nearest to the centre are *beau*, *grand* and *mauvais* which are generally considered as universal semantic features. Let's move away from the centre. There are now many semantic branches more or less long growing out from the central core. These branches are very dense near the centre and then go in all the directions becoming sparer and sparer. They are homogenous from a semantic viewpoint. Each branch only contains one sort of adjectival meanings: relationnel, qualificative, or intensive. In the outlying areas we find the meanings with a rich semantic contain like the relationnel meanings. Figure 2 shows a visualisation of the central sphere. We have also defined geometric tools to explore more precisely the different branches. We can't give all the details here[5].

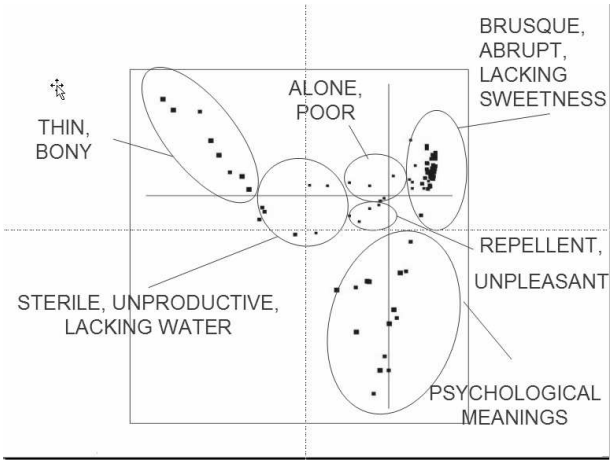

Fig. 1. Meaning's zones in the local semantic space associated with the French adjective *sec*

[4] Let $u_1, u_2 \ldots u_n$ denote the adjectives, and $c_1, c_2 \ldots c_p$ the cliques, the clique c_k corresponds to a point whose coordinate relatively at u_j is x_{ki} and $x_{ki} = 1$ if $u_i \in c_k$ and $x_{ki} = 0$ si $u_i \notin c_k$

[5] see F. Venant (06)

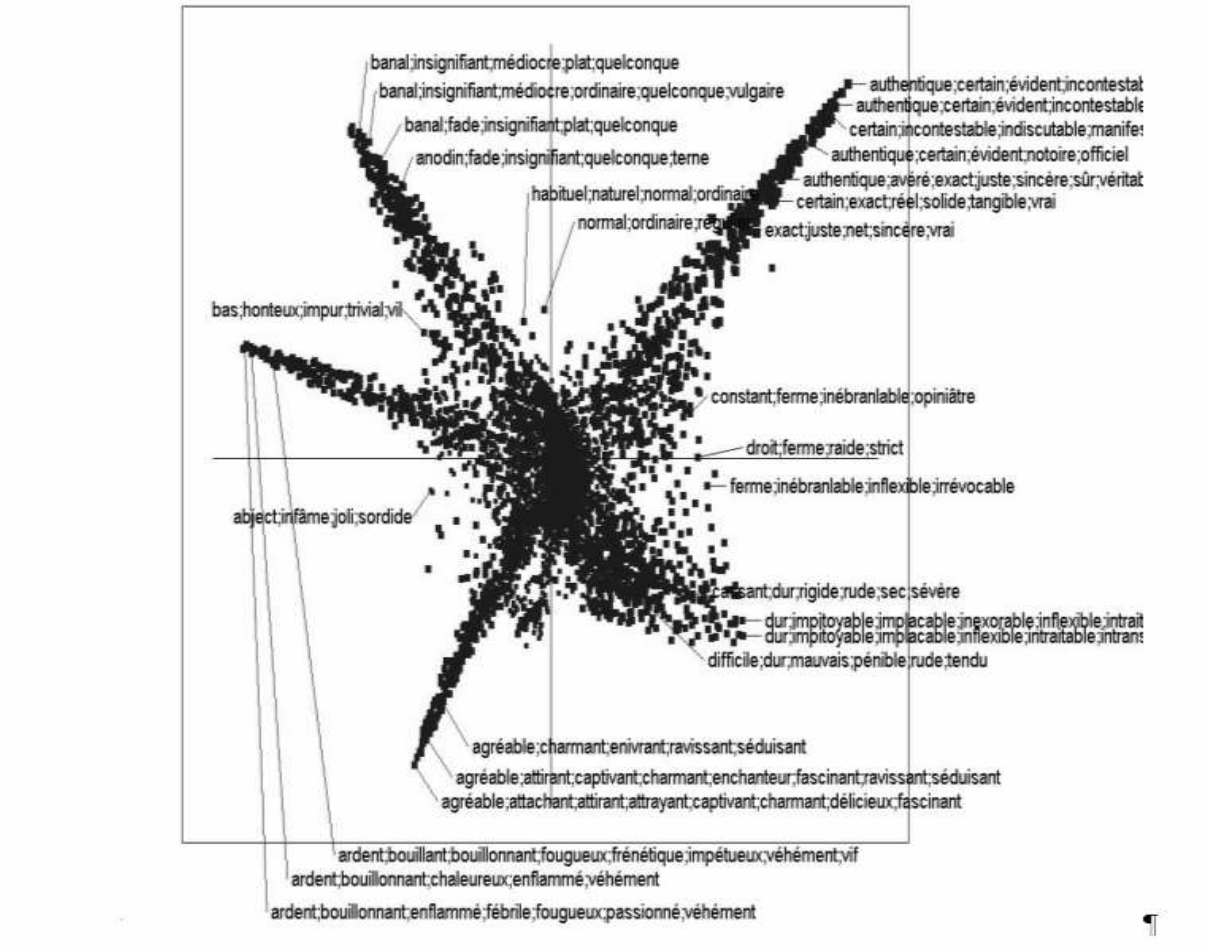

Fig 2, a global visualization of the central sphere

4. CONCLUSION

Our basic result is that the kind of visualization presented here displays the structure of the adjectival graph. From now we know that it likes a galaxy with a very dense central core. This galaxy is organized by the three kind of adjectival uses we drew out from the literature. This model accounts for theoretical studies on French adjectives. It also brings a new interesting light on the organisation of the French adjectival lexicon. We have then at our disposal tools to explore the structure of small-world graphs. This work is still in progress but it strengthens our belief in the similarity between graphs of lexical and social relationships. Our tools are independent of the nature of the relation modelized by the graph. As we said, the construction of 'semantic spaces' is not only interesting for lexical systems: it can prove very valuable for other 'semantic' graphs like the Web, as well as social graphs where the involved relationship depends more on conceptual factors than on geographical ones.

REFERENCES

[1] Barabási A. L., Jeong H., Neda Z., Ravasz E., Schubert A., and Vicsek T., 2002. Evolution of the social network of scientific collaboration, *Physica A*, 311(3-4):590-614

[2] Gaume B., Venant F., Victorri B., 2006. Langues, Hierarchy in lexical organisation of natural languages, *Springer Denise Pumain Hierarchy in Natural and social Sciences Methodos Series,* 121-142

[3] Milgram S., 1967. The small world problem, *Psychol. Today*, 2:60-67

[4] Newman M. E. J., 2001. The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA, 98:404-409

[5] Ploux S., Victorri B., 1998. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39(1):161-182

[6] Ravasz E., Barabási A.L, 2003. Hierarchical Organization in Complex Networks, *Phys. Rev. E*, 67, 026112

[7] Venant F., 2006. Représentation et calcul dynamique du sens: exploration du lexique adjectival du français., Thèse de l'Ecole des hautes études en sciences sociales - EHESS PARIS

[8] Watts D.J., Strogatz S.H., 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440-442